Highly precise protein-protein interaction prediction based on consensus between template-based and *de novo* docking methods

Masahito Ohue[1, 2], Yuri Matsuzaki[1], Takehiro Shimoda[1],
Takashi Ishida[1], Yutaka Akiyama[1§]

[1] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1-W8-76 Ookayama, Meguro-ku, Tokyo 152-8550, Japan.
[2] Research Fellow of the Japan Society for the Promotion of Science
[§] Corresponding author

Email addresses:
MO: ohue@bi.cs.titech.ac.jp
YM: y_matsuzaki@bi.cs.titech.ac.jp
TS: shimoda@bi.cs.titech.ac.jp
TI: t.ishida@bi.cs.titech.ac.jp
YA: akiyama@cs.titech.ac.jp

Abstract

Background

Elucidation of protein-protein interaction (PPI) networks is important for understanding disease mechanisms and for drug discovery. Tertiary-structure-based *in silico* PPI prediction methods have been developed with two typical approaches: a method based on template matching with known protein structures and a method based on *de novo* protein docking. However, the template-based method has a narrow applicable range because of its use of template information, and the *de novo* docking based method does not have good prediction performance. In addition, both these *in silico* prediction methods have insufficient precision and would therefore require validation of the predicted PPIs by biological experiments, leading to considerable expenditure; therefore, PPI prediction methods with greater precision are needed.

Results

We have proposed a new structure-based PPI prediction method by combining template-based prediction and *de novo* docking prediction. When we applied the method to the human apoptosis signaling pathway, we obtained a precision value of 0.333, which is higher than that achieved using conventional methods (0.231 for PRISM, a template-based method, and 0.145 for MEGADOCK, a non-template-based method), while maintaining an F-measure value (0.285) comparable to that obtained using conventional methods (0.296 for PRISM, and 0.220 for MEGADOCK).

Conclusions

Our consensus method successfully predicted a PPI network with greater precision than conventional template/non-template methods, which may thus reduce the cost of validation by laboratory experiments for confirming novel PPIs from predicted PPIs. Therefore, our method may serve as an aid for promoting interactome analysis.

Introduction

Elucidation of regulatory relationships among the tens of thousands of protein species that function in a human cell is crucial for understanding the mechanisms underlying diseases and for the development of medicines [1]. Predicting protein-protein interaction (PPI) networks at the genome scale is one of the main topics in systems biology.

The methods used for PPI network prediction include primary-structure-based searching [2, 3], evolutionary information-based methods [4], and tertiary-structure-based methods [5, 6]. Tertiary-structure-based methods are attracting attention because predicted protein complex structures are obtained and because these methods do not depend on homologous proteins. There are two typical approaches: a method based on template matching with known protein structures and another method based on *de novo* protein docking. The template-based method is based on the hypothesis that known complex structures or interface architectures can be used to model the complex formed between two target proteins. The hypothesis is logical, and this method provides good prediction performance when complex structural information is available as a template; however, if the template structure information is not available, performance is poor. In addition, because the interface architecture is not always similar for similar interactions, the template-based method has a narrow applicable range. In contrast, the *de novo* docking based method has a wide applicable range because it uses only tertiary structural information. However, because advantage of existing template information is not utilized, the prediction performance is poor.

Tuncbag *et al.* developed a template-based PPI prediction method called PRISM [5], which is based on information regarding the interaction surface of crystalline complex structures. PRISM has been applied for predicting PPIs in a human apoptosis pathway [7] and a p53-protein-related pathway [8] and has contributed to the understanding of the structural mechanisms underlying some types of signal transduction. Ohue *et al.* developed a PPI prediction method called MEGADOCK [6] and Wass *et al.* developed a method [9] based on protein-protein docking without interaction surface information. MEGADOCK has been applied for PPI prediction for a bacterial chemotaxis pathway [6] and has contributed to the identification of protein pairs that may interact.

However, the prediction results of both template-based and *de novo* docking-based methods in the studies contained many false-positive predictions. PRISM obtained a precision value of 0.231 when applied to a human apoptosis pathway that consisted of 57 proteins, which was higher than the precision obtained with random prediction (precision value of 0.086), and MEGADOCK obtained a precision value of 0.400 when applied to a bacterial chemotaxis pathway that consisted of 13 proteins, which was higher than the precision obtained with random prediction (precision value of 0.253). To identify new PPIs, the prediction results need to be validated using biological experiments. For this purpose, obtaining a low number of predicted interaction candidates with high reliability is more important than obtaining a high number of predictions with low reliability. Thus, this paper aims to improve the method used to obtain PPI predictions.

In this study, we combined two different PPI prediction methods to improve the precision of PPI prediction. Because PRISM is a template-based method, its prediction accuracy depends on the template dataset prepared. Only PPIs whose interaction surface structures are conserved are expected to be predicted. In contrast, MEGADOCK is a non-template-based method (also called *de novo* prediction), which has the demerit of generating false-positives for which no similar structures are seen in known databases; however, in situations where template structures are not present in databases, MEGADOCK can still predict PPIs. This qualitative difference between the two methods typically makes their output different. Thus, the combination of both prediction methods may improve prediction accuracy, as the intersection set (AND set) of both results may contain a few false positives; this improvement in accuracy would also contribute to improvement in the prediction reliability provided by the use of just one method.

Such an approach is called a “meta” approach. Meta approaches have already been used in the field of protein tertiary structure prediction [10], and critical experiments have demonstrated improved performance of meta predictors when compared with the individual methods used in the meta predictors. The meta approach has also provided favorable results in protein domain prediction [11] and the prediction of disordered regions in proteins [12]. We have therefore proposed a new PPI prediction method based on consensus between template-based and *de novo* docking methods. Generally, a meta prediction method may have low applicability because a meta approach requires applicable conditions of each single method. However, in this study, if structural information is available, *de novo* docking method is always applicable with or without template information. Thus, the applicability of the consensus method is not narrower than that of a template-based method.

Materials and Methods

Template-based PPI prediction

We used PRISM for template-based PPI prediction. PRISM uses two input datasets: the template set and the target set. The template set consists of interfaces extracted from protein pairs that are known to interact. The target set consists of protein chains whose interactions need to be predicted. The two sides of a template interface are compared with the surfaces of two target monomers by structural alignment. If regions of the target surfaces are similar to the complementary sides of the template interface, then these two targets are predicted to interact with each other through the template interface architecture.

The prediction algorithm consists of four steps: (1) interacting surface residues of target chains are extracted using Naccess [13]; (2) complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt [14]; (3) the structural alignment results are filtered according to threshold values and the resulting set of target surfaces is transformed onto the corresponding template interfaces to form a complex; and (4) the FiberDock [15] algorithm is used to refine the interactions to introduce flexibility, resolve steric clashes of side chains, compute the global energy of the complex, and rank the solutions according to their energies. This prediction protocol has been described in detail in a previous study [5].

PPI prediction based on the *de novo* docking method

For *de novo* protein docking-based PPI prediction, we used MEGADOCK version 2.6.2 [6]. MEGADOCK does not require template structures for prediction. The PPI prediction scheme used in this study consists of two steps. First, we conducted rigid-body docking calculations based on a simplified energy function considering shape complimentarity, electrostatics and hydrophobic interactions on all the possible binary combinations of proteins in the target set. Using this process, we obtained a group of high-scoring docking complexes for each pair of proteins. Next, we applied ZRANK [16] to the predicted complex structures for more advanced binding energy calculation and re-ranked the docking results based on ZRANK energy scores. The deviation of the selected docking scores from the score distribution of high-ranked complexes was determined as a standardized score (Z-score) and was used to assess possible interactions. This prediction protocol has been described in previous studies [17, 18]. Potential complexes that had no other high-scoring interactions nearby were rejected using structural differences. Thus, we considered likely binding pairs that had at least one populated area of high-scoring structures, one of which may be the true binding site.

Consensus prediction method

In this study, we proposed a new meta-prediction method by evaluating the consensus between both previously used prediction methods. The proposed method consists of two steps: (1) prediction from the same target set by PRISM and MEGADOCK and (2) considering that the method predicts the target protein pair interacts only when both PRISM and MEGADOCK predict the target protein pair interacts. Although some true-positives may be dropped by this method, the remaining predicted pairs are expected to have higher reliability because of the consensus between two prediction methods that have different characteristics.

Dataset

In this study, we focused on the human apoptosis signaling pathway previously analyzed by PRISM because our prediction results can thus be compared directly to the results of the previous study. PRISM and MEGADOCK are based on three-dimensional protein structures and therefore can only be applied to proteins whose tertiary structures are available. Therefore, we searched among proteins that were involved in the human apoptosis pathway and were present in the Protein Data Bank (PDB) (accessed at 28th July, 2012). We selected several proteins that had the highest resolution for the structural group that had high sequence similarity (>0.9) with the other proteins in the dataset [7]. After filtering according to resolution and sequence similarity, we obtained 158 PDB structures that corresponded to 57 proteins in the human apoptosis pathway described in KEGG (KEGG pathway ID: *hsa04210*) [19]. The PDB IDs in this structure dataset were the same as those used by Acuner Ozbabacan *et al* [7]. Table 1 shows the list of PDB IDs and chains of this dataset.

Known PPIs were collected from the STRING database [20]. We used only experimental data in the literature obtained from STRING with a confidence score >0.5. The number of known PPIs was 137. Because the database does not contain existing self-interactions, we did not predict self-interactions. Thus, the number of target pairs was ${}_{57}\mathbf{C}_2 = 1{,}596$.

Evaluation of prediction performance

Here, we have defined #TP, #FP, #FN, #TN, precision, recall, and the F-measure, which we used to evaluate the prediction results: #TP is the number of predicted PPIs that were also found in STRING (true positive), #FP is the number of predicted PPIs that were not in STRING (false positive), #FN is the number of PPIs not predicted by the system even though the pair was found to interact in STRING (false negative), and #TN is the number of negative predictions that were also not found in STRING (true negative). Precision, recall, and the F-measure are represented as follows:

$$\text{precision} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}}, \quad \text{recall} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}, \quad \text{F-measure} = \frac{2 \cdot \#\text{TP}}{2 \cdot \#\text{TP} + \#\text{FP} + \#\text{FN}},$$

where the F-measure is the harmonic mean of precision and recall. To identify new PPIs in biological experiments after *in silico* screening, precision is more important than recall.

Results and Discussion

Comparison of template- and non-template-based methods

Fig. 1(a) and (b) show the prediction results for PRISM and MEGADOCK, respectively, as applied to a human apoptosis pathway. The threshold used for MEGADOCK prediction yielded the best value of the F-measure for this dataset. The diagonal line (black cells) in Fig. 1 indicates self-interactions that were not considered as prediction targets. As shown in Fig. 1, PRISM was performed with fewer FPs than MEGADOCK. Table 2 shows the evaluation of prediction results. With MEGADOCK, we obtained a lower value of precision and a higher value of Recall relative to PRISM. When evaluating the F-measure as a measure of overall performance, MEGADOCK showed lower values than PRISM. Predictions by MEGADOCK contained more false positives because, in contrast to PRISM, MEGADOCK does not restrict interface structures to those found in template structures. In contrast, PRISM obtained lower recall values than MEGADOCK because it only searched interactions whose interface structures could be found in the template set.

Consensus prediction

Fig. 2 shows the Venn diagram of the number of TPs and FPs of the results of PRISM and MEGADOCK. A large difference was observed in the results obtained by the two methods. Thus, combining the prediction results of PRISM and MEGADOCK may provide better performance in PPI prediction.

Fig. 1(c) shows the prediction obtained on consensus between PRISM (a) and MEGADOCK (b); notably, the number of FP samples greatly decreased. The first row of Table 2 shows that the consensus method obtained an F-measure value of 0.285, which was comparable to the PRISM result (F-measure = 0.296). The consensus prediction indicated a higher value of precision for the consensus method (0.333) than for PRISM (0.231). The consensus method yielded the highest precision value in the method shown in Table 2. This method is useful when validating unknown PPI predictions using biological experiments. In contrast, OR prediction demonstrated high recall (Table 2). Thus, the OR method will be useful when prediction with high sensitivity, e.g., in the initial construction of the draft PPI network from the relevant proteins, is required.

In this study, we have used a fixed threshold value for MEGADOCK. The threshold value obtained the best F-measure value for the target dataset. Fig. 3 shows a plot of precision vs. F-measure for prediction results with various threshold values for MEGADOCK. Fig. 3 also plots the performance of the consensus method with various threshold values for MEGADOCK prediction while the threshold value for PRISM prediction was fixed. When the threshold value was changed in MEGADOCK, the plotted values remained in the region of low precision (0.0–0.2), and lower F-measure values were observed in the region of higher precision because of the decreased recall value. The consensus prediction method maintained a stable F-measure value when the value of precision was approximately 0.2–0.3, although its F-measure performance in the high-precision region (>0.4) was inferior to that of MEGADOCK. In this region, the consensus prediction provides a better precision value than PRISM, while maintaining the same F-measure level. Moreover, Fig. 3 clearly shows that the performance obtained by using the consensus method is better over a wide range of threshold values than the prediction obtained using only MEGADOCK.

Conclusions

In this study, we propose a new PPI network prediction method based on the consensus between template-based prediction and non-template-based prediction. The consensus method successfully predicted the PPI network more accurately than the conventional single template/non-template method. Because such precise prediction can reduce biological screening costs, it will serve to promote interactome analysis. For further improvement of prediction performance, it is necessary to further improve the combination of the two techniques, i.e., by using a strategy other than taking a simple AND/OR consensus. For example, biological information such as biochemical function and subcellular localization information could be used.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MO developed the consensus interaction prediction method, designed the human apoptosis pathway problem and wrote the manuscript. MO and YM performed the computational experiments and validated the results. TS performed the PRISM experiments. TI assisted with the method design. YA supervised and directed the entire study. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by a Grant-in-Aid for JSPS Fellows (238750), a Grant-in-Aid for Research and Development of The Next-Generation Integrated Life Simulation Software, and by the Education Academy of Computational Life Sciences, all of which were from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). Parts of the results were obtained by early access to the K computer at the RIKEN Advanced Institute for Computational Science.

References

1. Wass MN, David A, Sternberg MJE: **Challenges for the prediction of macromolecular interactions**. *Curr. Opin. Struct. Biol.* 2011, **21**:382–390

2. Higurashi M, Ishida T, Kinoshita K: **Identification of transient hub proteins and the possible structural basis for their multiple interactions**. *Protein Sci.* 2008, **17**:72–78

3. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H: **Predicting protein-protein interactions based only on sequences information**. *Proc. Natl. Acad. Sci. U. S. A.* 2007, **104**:4337–4341

4. Valencia A, Pazos F: **Prediction of protein-protein interactions from evolutionary information**. *Structural Bioinformatics, Second Edition*, Wiley and Sons: New York 2009, 617–634

5. Tuncbag N, Gursoy A, Nussinov R, Keskin O: **Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM**. *Nature Protocols* 2011, **6**:1341–1354

6. Ohue M, Matsuzaki Y, Ishida T, Akiyama Y: **Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model : an application to interaction pathway analysis**. *Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science* 2012, **7632**:178–187

7. Acuner Ozbabacan SE, Keskin O, Nussinov R, Gursoy A: **Enriching the human apoptosis pathway by predicting the structures of protein-protein complexes**. *J. Struct. Biol.* 2012, **179**:338–346

8. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R: **Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example**. *Mol. Biosyst.* 2009, **5**:1770–1778

9. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A: **Towards the prediction of protein interaction partners using physical docking**. *Mol. Syst. Biol.* 2011, **7**:469

10. Zhou H, Pandit SB, Skolnick J: **Performance of the Pro-sp3-TASSER server in CASP8**. *Proteins* 2009, **77**:123–127

11. Saini HK, Fischer D: **Meta-DP: domain prediction meta-server**. *Bioinformatics* 2005, **21**:2917–2920

12. Ishida T, Kinoshita K: **Prediction of disordered regions in proteins based on the meta approach**. *Bioinformatics* 2008, **24**:1344–1348.

13. Hubbard SJ, Thornton JM: **Naccess**. Department of Biochemistry and Molecular Biology, University College London, 1993

14. Shatsky M, Nussinov R, Wolfson HJ: **A method for simultaneous alignment of multiple protein structures**. *Proteins* 2004, **56**:143–156

15. Mashiach E, Nussinov R, Wolfson HJ: **FiberDock: Flexible induced-fit backbone refinement in molecular docking**. *Proteins* 2010, **78**:1503–1519

16. Pierce B, Weng Z: **ZRANK: reranking protein docking predictions with an optimized energy function**. *Proteins* 2007, **67**:1078–1086

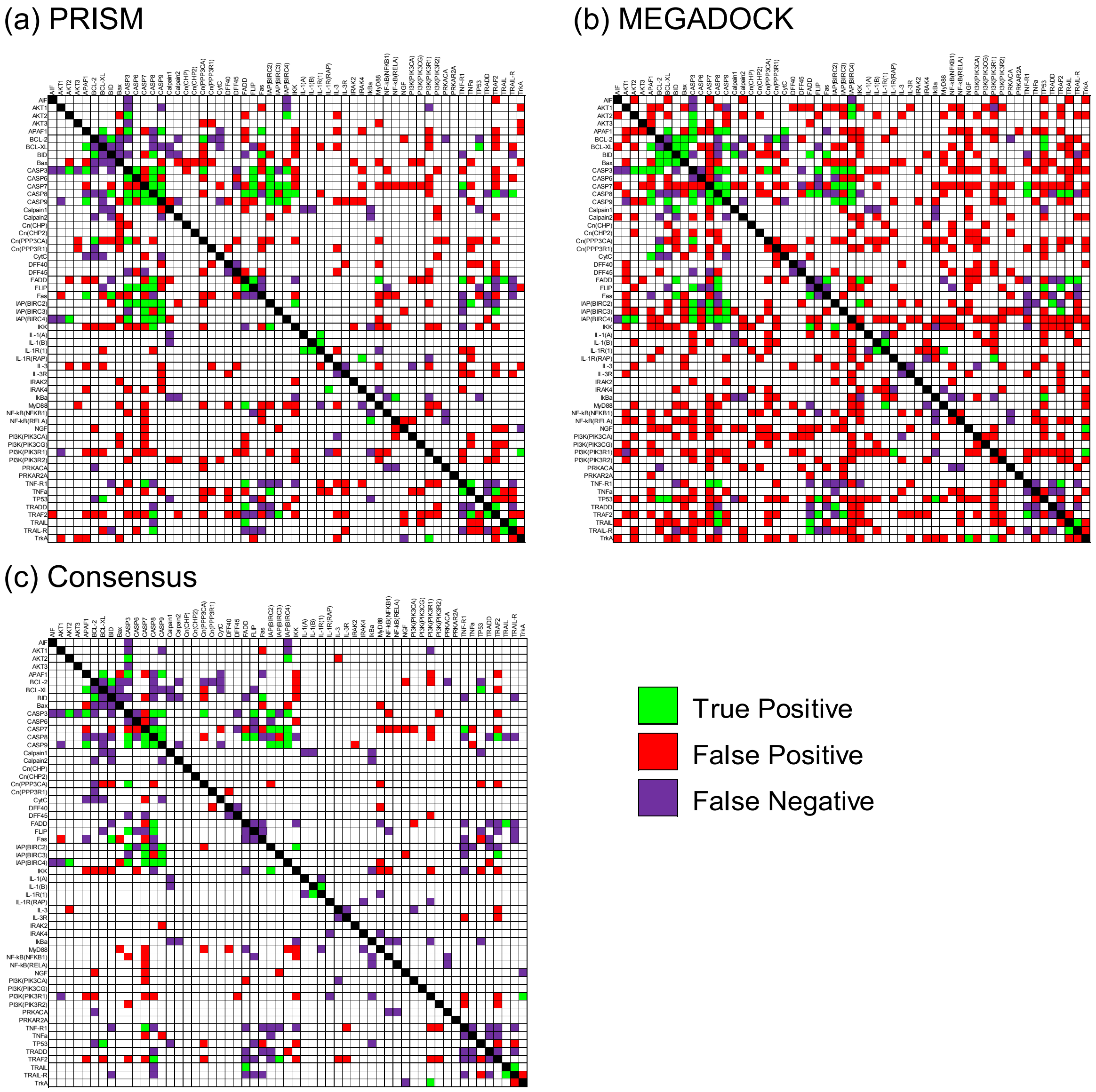
17. Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y: ***In silico* screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis**. *J. Bioinform. Comput. Biol.* 2009, **7**:991–1012

18. Ohue M, Matsuzaki Y, Akiyama Y: **Docking-calculation-based method for predicting protein-RNA interactions**. *Genome Informatics* 2011, **25**:25–39

19. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res.* 2000, **28**: 27–30.

20. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Res.* 2011, **39**: D561–568.

Figures


(a) PRISM

(b) MEGADOCK

(c) Consensus

Figure 1 - Apoptosis prediction by the (a) PRISM, (b) MEGADOCK, and (c) consensus methods.

The green cells are true positives, the red cells are false positives, and the purple cells are false negatives. The diagonal cells (black cells) have no PPI information in the STRING database and are excluded from the prediction targets.

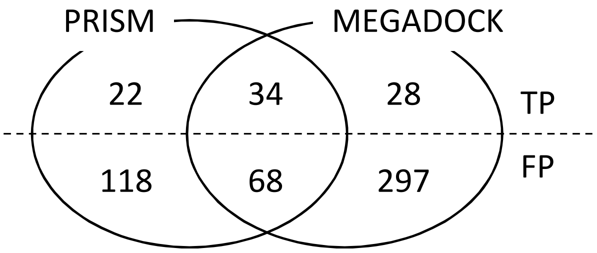


Figure 2 - Venn diagram of apoptosis pathway prediction results

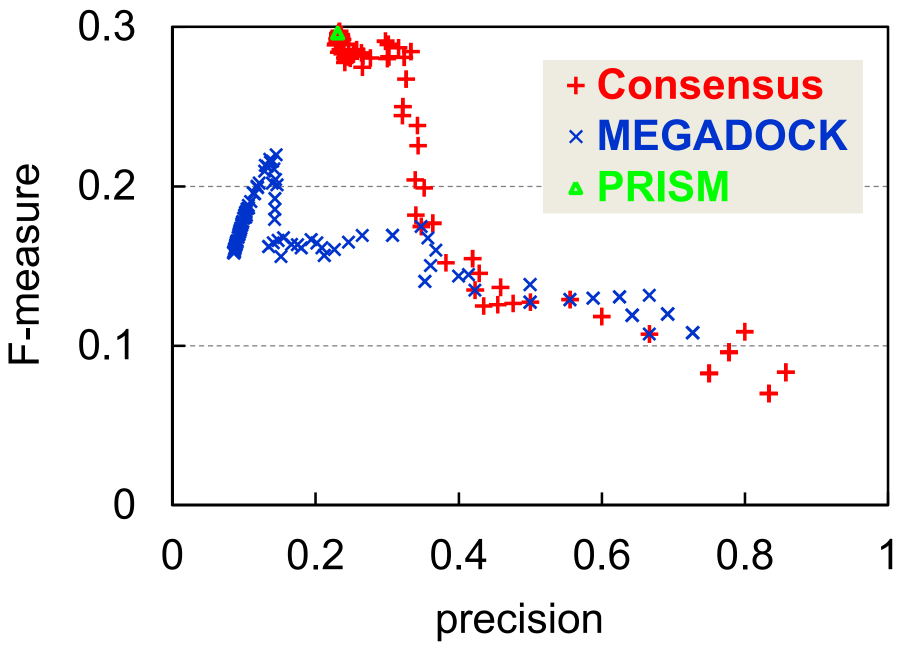


Figure 3 - F-measure vs. precision for predictions when the threshold parameter of MEGADOCK is changed in the apoptosis pathway

The green triangle indicates the result of PRISM prediction (Table 2).

Tables

Table 1 - Protein and PDB ID list of human apoptosis pathway dataset

Protein Name	PDB ID (_Chain)					
AIF	1M6I_A					
AKT1	1UNQ_A	3CQW_A	3O96_A			
AKT2	1MRV_A	1O6K_A	1O6L_A	1P6S_A		
AKT3	2X18_A					
APAF1	1CY5_A	1Z6T_A	2YGS_A	3IZA_A	3YGS_C	
BCL-2	2W3L_A	2XA0_A				
BCL-XL	2B48_A	3FDL_A				
BID	2BID_A	2KBW_B				
Bax	1F16_A	2G5B_I	2XA0_C	3PK1_B		
CASP3	1RHQ_A	1RHQ_B	2DKO_A	2DKO_B	2J32_A	
CASP6	2WDP_A					
CASP7	1F1J_A	1I4O_A	1I51_A	1I51_B	2QL9_A	2QL9_B
CASP8	1QTN_A	1QTN_B	2FUN_B	3H11_B		
CASP9	1JXQ_A	1NW9_B	3D9T_C	3YGS_P		
Calpain1	1ZCM_A					
Calpain2	1KFU_L	2NQA_A				
Cn(CHP)	2E30_A					
Cn(CHP2)	2BEC_A					
Cn(PPP3CA)	1AUI_A	1MF8_A	2R28_C	3LL8_A		
Cn(PPP3R1)	1AUI_B	1MF8_B	3LL8_B			
CytC	1J3S_A					
DFF40	1IBX_A					

DFF45	1IBX_B	1IYR_A				
FADD	1A1W_A	2GF5_A	3EZQ_B			
FLIP	3H11_A					
Fas	3EWT_E	3EZQ_A				
IAP(BIRC2)	3D9T_A	3M1D_A	3MUP_A			
IAP(BIRC3)	2UVL_A	3EB5_A	3EB6_A	3M0A_D	3M0D_D	
IAP(BIRC4)	1G73_C	1I4O_C	1I51_E	1NW9_A	2ECG_A	
	2KNA_A	2POI_A	3CM7_C			
IκBα	1IKN_D	1NFI_E				
IKK	2JVX_A	3BRT_B	3BRV_B	3CL3_D	3FX0_A	
IL-1(A)	2ILA_A					
IL-1(B)	1ITB_A	2NVH_A	3O4O_A			
IL-1R(1)	1ITB_B					
IL-1R(RAP)	3O4O_B					
IL-3	1JLI_A					
IL-3R	1EGJ_A					
IRAK2	3MOP_K					
IRAK4	2NRU_A	3MOP_G				
MyD88	2JS7_A	3MOP_A				
NF-κB(NFKB1)	1IKN_C	1NFI_B	1SVC_P	2DBF_A		
NF-κB(RELA)	1IKN_A	1NFI_A				
NGF	1WWW_V	2IFG_E				
PI3K(PIK3CA)	2ENQ_A	2V1Y_A	3HHM_A			
PI3K(PIK3CG)	1E8Y_A					
PI3K(PIK3R1)	1A0N_A	1H9O_A	1PBW_A	2IUG_A	2V1Y_B	
	3HHM_B	3I5R_A				
PI3K(PIK3R2)	2KT1_A	2XS6_A	3MTT_A			
PRKACA	3AGM_A					
PRKAR2A	2IZX_A					
TNFα	1A8M_A	4TSV_A				
TNF-R1	1EXT_A	1ICH_A				
TP53	1AIE_A	1OLG_A	1XQH_B	1YC5_B	2B3G_B	2FOO_B
	2GS0_B	2K8F_B	2VUK_A	3D06_A	3DAB_B	3LW1_P
TRADD	1F3V_A					
TRAF2	1CZZ_A	1D00_A	1F3V_B	3KNV_A	3M0A_A	3M0D_A
TRAIL	1D4V_B	1DG6_A	1DU3_D			
TRAIL-R	1D4V_A	1DU3_A				
TrkA	1HE7_A	1SHC_B	1WWW_X	2IFG_A		

Table 2 - Accuracy of human apoptosis pathway prediction

Method	#TP	#FP	#FN	#TN	precision	recall	F-measure
Consensus(AND)	34	68	103	1,391	0.333	0.248	0.285
OR	84	483	53	976	0.148	0.613	0.239
MEGADOCK	62	365	75	1,094	0.145	0.453	0.220
PRISM	56	186	81	1,273	0.231	0.409	0.296